基于知识图谱的学科单选题考点提取研究 *

刘一然，骆力明

(首都师范大学 信息工程学院，北京 100048)

摘　要： 在人工智能快速发展的今天，智能教育逐渐成为一大研究热点。在自然语言处理方面对智能教育中智慧学习的探究，提出根据知识图谱和学科规则确定单选题考点，主要介绍知识图谱的构建和单选题考点的提取这两方面内容。通过建立一个开放性的知识图谱，不断实现学科知识的扩充。为提取考点，首先将单选题分类、分词以及替换相似词，然后通过检索图谱得到单选题的候选考点集，最后通过学科规则定位知识点及其所属章节，便于学生有针对性地复习教材知识。在所收集的C++试题集上的实验结果表明，通过知识图谱和规则可较为准确地提取出试题考点。

关键词： 知识图谱；单选题；知识点；学科规则

中图分类号： TP391　　doi: 10.3969/j.issn.1001-3695.2017.12.0839

Research on extraction of subject single knowledge points based on knowledge map

Liu Yiran, Luo Liming

(*Information Engineering School*, *University of Capital Normal*, *Beijing* 100048, *China*)

Abstract: At present, artificial intelligence develops rapidly, intelligent education has gradually become a hot research. This paper is a study of intelligent learning in intelligent education in the aspect of natural language processing. This paper proposed to determine the single choice knowledge according to the knowledge map and discipline rules. It mainly introduced the construction of knowledge map and the extraction of the single choice knowledge, through the establishment of an open knowledge map, expand the subject knowledge. In order to extract the test knowledge, first, classification of single choice, word segmentation and replacement of similar words, then, the candidate knowledge set of single choice can be obtained by searching the knowledge map, finally, through the discipline rules to locate the knowledge points and their chapters, help students to targeted review the knowledge of teaching materials. The experimental results on the collected C++ test set show that the test points can be accurately extracted by knowledge map and rules.

Key words: knowledge map; single choice; knowledge point; discipline rules

0　引言

近年来，人工智能（artificial intelligence）对社会各领域的影响正逐渐加深，现如今，它已经渗入到了金融、医疗、农业、国防和环保等领域。人工智能技术在金融行业已被广泛应用。以国内著名的电子商务公司——阿里巴巴为例，利用人工智能技术，在客户服务、征信、智能投顾、保险、互联网小贷等多个领域进行创新和应用[1]；在医疗领域，人工智能技术已被初步应用在智能诊断、智能治疗、日常化护理、人性化医疗等方面[2]；在智慧农业方面，利用数据采集技术、无线通信技术和计算机技术对大型塑料温室进行智能化监测、管理和控制，提高农业信息化水平[3]。

人工智能与教育也在不断地融合与发展，专家系统、智能教学系统（ITS）、智能决策支持系统、智能计算机辅助教学（CAI）系统发展迅速[4]，出现了一系列的教育智能产品。例如，讯飞畅言智慧校园不仅可以帮助教师有针对性地安排教学进度和内容，而且可以根据学生综合素质发展情况进行个性化指导和差异化教学；基于物联网技术的 HappyClass 智慧课堂系统，可实现教师与学生双向实时“一对多”互动教学，VR、AR 情境教学等；基于现实世界中的各种校园信息，腾业智慧校园建立了一种虚拟教育环境。

自然语言处理是人工智能领域的一个重要分支，它的目的是让计算机理解人类的自然语言，从而实现用自然语言与计算机进行交流[5]。自然语言处理技术在教育领域的应用越来越广泛，有研究者将其概括为以下四方面[6]：文本的分析与知识管理，如作文自动评价等；人工系统的自然语言界面，如智能问

收稿日期：2017-12-08；修回日期：2018-01-24　**基金项目：** 国家自然科学基金面上项目（61672361）；北京成像技术高精尖创新中心资助项目（BAICIT-2016004）

作者简介： 刘一然（1993-），女，硕士研究生，主要研究方向为智能教育（lyr_cnu@126.com）；骆力明（1963-），男，教授，主要研究方向为智能教育、软件系统实现.

答系统等；语料库在教育中的应用，如基于语料库的数据挖掘工具等；面向语言教学研究的应用，如计算机辅助语言教学等[7]。

本研究着眼于教学中的考试环节，在学科考试中，单选题有着题量大、考查面广等特点，快速提取单选题考点，可以帮助学生及时了解学科知识的薄弱环节，从而有的放矢地填补缺漏；把题目考点定位到教材章节，可以帮助学生学习章节知识，故图谱知识点是按照教材章节组织的。

标注单选题的考点后，笔者发现根据考点分布位置可将单选题分为两类：一类是通过分析题目便可确定具体考点的单选题；另一类是通过分析题目和选项才能决定具体考点的单选题。分类后，使用分词、同义词替换技术对文本进行处理，接着用题中词语匹配图谱中知识点的关键词，得到候选考点集，最后依据学科规则得到准确知识点。

1　系统整体架构

图 1 给出了基于知识图谱提取单选题考点的系统架构。逻辑上分为两层：语料处理层和检索知识点层。

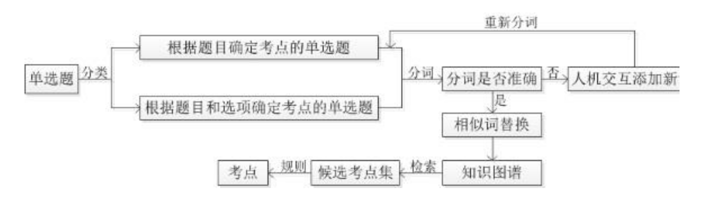


图 1　基于知识图谱提取单选题考点的系统架构

1.1　语料处理

这一过程可分为题目分类、语料分词以及相似词替换这三步。

1)题目分类。根据建立的模板对题目分类，从而决定后续知识点的提取方式：

①只对单选题题目进行分析。

②对整道试题（题目及其选项）进行分析。

2）语料分词。系统中的分词模块是开放式的，通过人机交互，用户可以向自定义词典中添加新词，从而提高分词准确率。

3）相似词替换。归纳总结试题以及课本中语义相同的词语，据此构造相似词典并利用该词典统一题中词汇为一常见表达，从而提高知识点检索的准确性。

上述三步骤的细节将在第 2 章中作详细描述。

1.2　检索知识点

此过程需学科知识图谱的参与，将上一流程得到的单选题词汇与知识图谱中储存的知识点的关键词进行匹配，匹配成功的知识点构成该题的候选考点集。该考点集由基本知识点和主题知识点构成，即，如果题目考查某一主题下的基本知识点，则匹配时会将该主题知识点一并提取出，而主题知识点是一个泛化的知识点，在能提取出基本知识点的情况下，它通常是不被需要的。

一种特殊的情况是候选考点集由多个基本知识点组成，有两种原因，一是因为题中存在主题知识点词汇及其多个基本知识点的描述，组合后可得到多个基本知识点；二是因为含义不同的考点可能由相同的领域关键词组成，若题目中存在这些关键词，则在匹配时会将这些含义不同的考点全部提出。

为得到准确考点，笔者针对学科知识点建立了学科规则。以上描述的具体细节将在第 3 章中介绍。

1.3　数据集介绍

1.4　学科教材

本研究力图使单选题考点对应到具体的教材章节，故图谱中的知识点是按照其所属章节组织的。在分析了章节标题及其对应内容后，笔者发现章节标题中的领域名词通常是一个主题知识点，从其具体的章节内容中可提取出该主题知识点下的许多具体知识点，故学科知识图谱的完整性一方面取决于领域专家对章节知识的总结是否全面，另一方面取决于所选取的教材内容是否完整。

本研究选取大学课程 C++作为实验对象，它既是一种被计算机专业学生广泛使用的编程语言，又是计算机二级考试中的一个考试科目，故其有着十分中要的教学地位。本实验选用电子工业出版社出版的由杜茂康等人编著的《C++面向对象程序设计（第 2 版）》这本教材，该书是高等学校工程创新型“十二五”规划的计算机教材，全书共十二章，本实验只研究介绍标准 C++面向对象程序设计技术的前九章[8]。各章节的主题知识点及所对应的具体知识点的数量统计结果如表 1 所示。

表 1　教材各章节主题知识点及具体知识点数目统计结果

教材章节	主题知识点	具体知识点
第 1 章 C++与面向对象程序设计概述	11	34
第 2 章　C++基础	33	99
第 3 章　类与对象	25	82
第 4 章　继承	21	53
第 5 章　多态性	8	20
第 6 章　运算符重载	14	45
第 7 张　模板与 STL	12	76
第 8 章　异常	11	21
第 9 章　文件与流	9	46

1.5　单选题库

目前，题库中共储存了 1 500 多道 C++单选题，笔者对每题都进行了考点标记。

1.5.1 单选题分类

观察标记考点后的 C++单选题，笔者发现单选题的考点在题中的位置分布是有一定规律的，一部分单选题的考点可通过分析其题目直接确定，一部分单选题的考点要根据其题目和选项共同确定，还有部分单选题的考点只分布在其选项中。在此，笔者将最后两种情况合为一种，即通过分析整道单选题确定考点位置。

为实现单选题自动分类，笔者观察到第二类单选题的题目中存在有特殊字符串，如“叙述正确的是”，“正确的说法是”等，这些字符串既可匹配题目中不包含任何考点的单选题，如“下列叙述正确的是”，又可匹配题目包含主题知识点而选项包含具体知识点的单选题，如“下列关于构造函数的叙述正确的是”。题库单选题分类结果见第 4 章。

1.5.2 单选题分词

汉语分词是处理中文语料必不可少的一个环节，其在自然语言处理中居于基础地位。目前较为常用的分词工具有语言技术平台（LTP）、NLPIR 汉语分词系统、开源自由的汉语言处理包（HanLP）以及 Python 自带的中文分词组件 jieba。

在实验前期，笔者曾尝试使用以上这些分词器对中文语料进行分词，后期考虑到实验是在 Python 环境中进行的，所以优先考虑使用 jieba 分词组件。

该分词组件的优点在于安装简单、调用方便以及支持添加自定义词典。不同的学科有其特定的学科词语，当 jieba 词库中未包含这些词语时，分词结果往往不尽人意，因此需要添加自定义词典，提高分词准确率。本系统中的分词模块是人机交互式的，在自动分词后，用户可判断结果中是否存在错分词。若存在，可以通过交互将新词添加到自定义词典中，如此便可按照用户需要准确分词。分词模块的执行流程如图 2 所示。

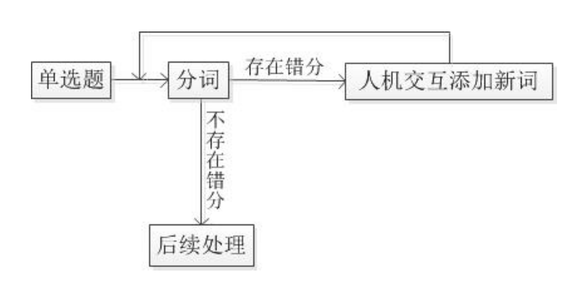


图 2　系统分词流程

1.5.3 相似词替换

由于汉语同义词较多且不同出题者对同义词语的使用有着不同程度的偏好，所以需统一分词后的词语表示，将意思表述相同的词语统一为一个最为常见的词，该词语必须和知识图谱中知识点的关键词用词一致，从而便于后续检索。

2　知识图谱

为将题中考点对应到教材章节，在构建知识图谱[9,11]时需归纳总结章节知识点，提取各个知识点的名称关键词和有关知识点描述的关键词。

2.1　构建知识图谱

将归纳总结出的知识点名称、知识点名称关键词、知识点概念描述的关键词以及该知识点所属章节储存在一个 json 文件中。选用 json 文件储存知识点的原因在于它是一种人类可读的、有层次结构且易于解析的文本数据交换格式。知识点在文件中的存储形式如下：

```
"chapter": "章节名称",
"knowledge":
[
{
"keywords": ["关键词 1","关键词 2",……],
"name": "知识点名称",
"description": ["关键词 1", "关键词 2",……]
},
……
]
```

其中，“keyword”中的关键词是从知识点名称中提取出的领域关键词；“description”中的关键词是知识点概念描述中的关键词，建立该字典的原因是：在学科考试中，单选题常常会考查某个知识点的基础概念，考查方式分为两种，一种是在题目中直接询问某个知识点的概念是什么，另一种考查方式则是在单选题的题目中给出某一知识点的概念，然后问考查的是哪个知识点。“description”就是针对第二种考查方式建立的。

为完善领域知识，本研究中的知识图谱是开放性的，通过交互，用户可不断添加新知识。

2.2　检索知识图谱

2.2.1 匹配关键词得到候选考点集

检索知识图谱是本研究中最为关键的一步，检索结果将直接决定能否成功提取单选题的考点。本文第章节介绍可将单选题按照其考点分布位置分为两类；相应地，在提取知识点时，也只需对题中考点的分布位置进行查找，这样不仅可以提高检索效率而且可以避免引入一些无关考点。具体步骤如下：

a）将试题按其题目考点分布进行分类。

试题集 A：通过分析题目便可确定考点的试题。

试题集 B：通过分析题目和选项从而确定考点的试题。

b)对两类试题集分别进行预处理。

试题集 A′：将试题集 A 中的试题的题目分词并对结果进行相似词替换。

试题集 B′：对试题集 B 中的试题进行分词操作并对每题的分词结果进行相似词替换。

c)用试题集 A′去检索知识图谱，得到试题的候选考点集。

遍历试题集 A′，用本次遍历的试题词语集与知识图谱中某一知识点的“keyword”中存储的关键词进行匹配，若能完全匹配，便可将知识点确认为该单选题的一个候选知识点，若匹配成功，则继续匹配下一知识点；若与“keyword”中的词语匹配失败，则会与该知识点的“description”中的关键词进行匹配，匹配成功的判断标准不变，若匹配失败，则继续匹配下一个知识点，直至与知识图谱中的最后一个知识点匹配结束。

d)用试题集 B′去检索知识图谱，得到每题的候选考点集。方法同 c），这里不再赘述。

2.2.2 设定规则得到准确考点

上述方式得到的考点并不准确，候选考点集中存在冗余考点和非题目考点的情况。

若单选题考查的是基本知识点，用该方法会将主题知识点也提取出来，即冗余考点。若某几个知识点其意思表述不同但关键词语相同，例如，“友元函数不是类的成员函数”和“一个类的所有成员函数都可以是另一个类的友元函数”，这两个考点都可提取出“类”、“友元函数”以及“成员函数”这三个词语，但两者表述的意思不同；若题中存在学科名词和几个通用的描述性词语，例如，在 C++单选题中，若题中存在“构造函数”“定义”“调用”这三个词语，便会匹配到两个考点，一个是“构造函数的定义”，一个是“构造函数的调用”，仅根据这三个词语的描述无法确定题目究竟考察的是哪个考点。通过设定规则，笔者消除了这两种情况带来的影响。方法如下：

a）删除冗余考点。

可比对候选考点的名称，若某一知识点的名称字符串是另一知识点名称字符串的子串，即可确定该知识点是主题知识点另一个知识点是具体知识点，此时应删去主题知识点。

b）通过分析词语在单选题中出现的先后顺序，区分具有相同关键词的知识点。

c）通过比较词语间的距离，确定学科关键词究竟该和哪个描述性词语结合。

①计算学科关键词和描述性词语在试题中所处的位置。

②分别计算学科关键词与各个描述性词语位置距离的绝对值。

③绝对值最小的距离所对应的描述性词语即为所求。

3 实验结果

3.1 试题分类及分词结果

通过特殊字符串的匹配将题目分为两类，具体结果见图 3。

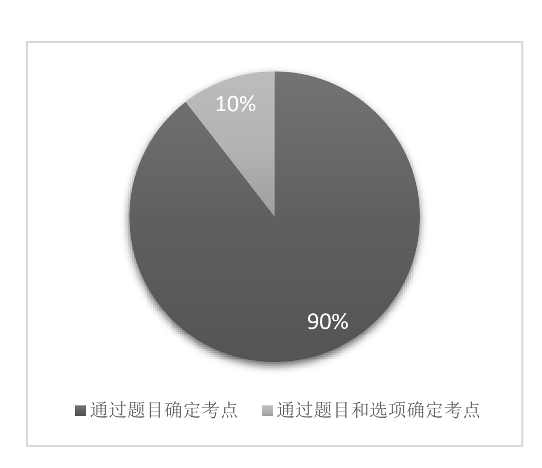


图 3 依据题中考点位置分布的单选题分类

从饼状图中可以清楚地看出本实验所搜集的单选题中有 90%的试题通过分析题目便可确定考点。

在语料处理时，比较了自动分词的准确率和添加自定义词典后的分词准确率。具体情况见图 4。

从图 4 可看出，自动分词的正确率只有 60%，这远不能满足后续分析的需求，添加自定义词典后，准确率提高到了 96%，极大地减弱了分词结果对考点提取准确率的影响。

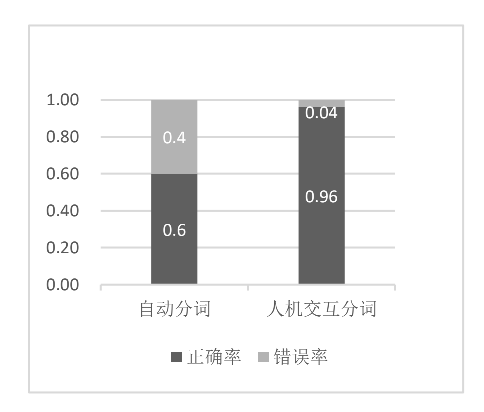


图 4 自动分词与添加自定义词典后的分词准确率对比图

3.2 匹配知识图谱结果

实验过程中，通过匹配题中关键词与知识图谱中知识点的关键词，得到单选题的候选考点集。具体情况见表 2。

表 2 候选考点集的组成情况

是否存在准确题目考点	是否存在冗余考点	是否存在非题目考点	题目所占比例
是	是	否	38%
是	是	是	13%
是	否	否	33%
是	否	是	16%

表 2 说明，通过检索知识图谱、直接匹配题目词语，能得到准确考点的题目数占总题目数的比重约为 33%，这一结果显然不能满足实际需要。针对这一现象，笔者设计学科规则，极大提高了准确率。具体情况见表 3。

表 3 匹配规则结果

能提取出准确考点	不能提取出准确考点
97%	3%

从表 3 结果可知，经过规则的处理，考点提取的准确率达到 97%。

分析通过规则匹配不能得到准确考点的题目，发现这类题目的特点往往是题中存在较多的学科名词，这些学科名词与通用的描述性词语组合，会得到较多的冗余考点和非题目考点，这会极大地影响最终结果。后期将尝试通过语义分析来提取这类试题的考点。

4 结束语

研究表明，通过检索学科知识图谱和匹配学科规则可准确提取学科单选题考点。本研究提出的方法有助于教师和学生对错题考点的分析，故可将其应用于学校教育。具体地，对于学生，错题考点其实就是他尚未完全掌握的学科知识，本研究提出的方法能快速且准确地将考点定位到教材章节，从而便于学生学习。对于教师，本方法可帮助他迅速提取学生错题考点，从而决定之后的题目讲解重点。

参考文献：

[1] 喻家骏. 人工智能在金融领域的应用 [J]. 电子技术与软件工程, 2016 (24): 158.

[2] 高奇琦, 吕俊延. 智能医疗：人工智能时代对公共卫生的机遇与挑战 [J]. 电子政务, 2017 (11): 11-19.

[3] Lu Huiguo, Li Congying, Jiang Juanping. Application of intelligence control in agriculture greenhouses [J]. Applied Mechanics and Materials, 2015, 719-720 (01): 293-297.

[4] 吴永和, 刘博文, 马晓玲. 构筑“人工智能+教育”的生态系统 [J]. 远程教育杂志, 2017, 35 (5): 27-39.

[5] 王海芳, 李峰. 人工智能应用于教育的新进展 [J]. 现代教育技术, 2008, 18 (13): 18-20.

[6] 王萌, 俞士汶, 朱学锋. 自然语言处理技术及其教育应用 [J]. 数学的实践与认识, 2015, 45 (20): 151-156.

[7] 闫志明, 唐夏夏, 秦旋, 等. 教育人工智能 (EAI) 的内涵、关键技术与应用趋势——美国《为人工智能的未来做好准备》和《国家人工智能研发战略规划》报告解析 [J]. 远程教育杂志, 2017 (1): 26-35.

[8] 杜茂康, 李昌兵, 曹慧英, 等. C+面向对象程序设计 [M]. 2 版. 北京: 电子工业出社, 2011: 1-252.

[9] 刘知远, 孙茂松, 林衍凯, 等. 知识表示学习研究进展 [J]. 计算机研究与发展, 2016, 53 (2): 247-261.

[10] 焦晓静, 王兰成. 知识图谱的概念辨析与学科定位研究 [J]. 图书情报工作, 2015 (15): 5-11.

[11] 胡泽文, 孙建军, 武夷山. 国内知识图谱应用研究综述 [J]. 图书情报工作, 2013, 57 (3): 131-137.